Sohn, Alexander; Klein, Nadja; Kneib, Thomas

Conference Paper

A new semiparamtetric approach to analysing Conditional Income Distributions

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Econometric Theory, No. C20-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Sohn, Alexander; Klein, Nadja; Kneib, Thomas (2014) : A new semiparamtetric approach to analysing Conditional Income Distributions, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Econometric Theory, No. C20-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/100630

A new semiparametric approach to analysing Conditional Income Distributions

Working Version 0.1

Alexander Sohn, Nadja Klein and Thomas Kneib

Abstract

In this paper we explore the application of Generalised Additive Models of Location, Scale and Shape for the analysis of conditional income distributions in Germany following the reunification. We find that conditional income distributions can generally be modelled using the three parameter Dagum distribution and our results hint at an even more pronounced effect of skill-biased technological change than can be observed by standard mean regression.

JEL-Classification:C13 , C21, D31, J31

1 Introduction

Following the publication of the fourth report on poverty and wealth (BfAS, 2013) a public debate on the extent and the nature of income inequality in Germany (re)erupted. While the dust hasn't fully settled yet, there is an increasing consensus that with some delay Germany is partially catching up with what Paul Krugman (2007) called the Great Divergence of incomes. One hypothesis put forward to explain this divergence is a skill-biased technological change tilting the labour demand in favour of those with higher educational attainment (Acemoglu, 2002). Various studies have analysed the changing impact of education on the German distribution of labour incomes (see among others Dustmann et al., 2009; Card et al., 2013). A second rather country-specific aspect which has received great attention in the literature is the impact of the persistent difference of former East and West Germany after reunification more than 20 years ago (Fuchs-Schündeln et al., 2010). While various authors have attempted to quantify the impact of the increasing wage-gap between the skill groups (Fitzenberger, 1999, e.g.) and others have found convergence of mean incomes between East and West (Vollmer et al., 2013) it remains true that "we know relative little about the determinants of residual inequality" (Acemoglu, 2002). It is the aim of this paper to put residual inequality, or rather of conditional income distributions (CIDs), in the centre of the analysis.[1] Specifically, we investigate conditional labour income distributions of men dependent on age, region and education for the years 1992 and 2010. We thus attempt to deviate from the dichotomy of economic analysis which either focusses on conditional means

We thank the DIW for the data and Marcus Grabka as well as Stefan Bach for their friendly support.
For correspondence please contact Alexander Sohn, Chair of Statistics, University of Göttingen, asohn@uni-goettingen.de

[1] We thus pursue the same aim as Chernozhukov et al. (2013) but by a different procedure.

(and seldomly variances) or the (aggregate) income distribution. For this purpose we introduce the class of Generalised Additive Models of Location, Scale and Shape (GAMLSS) first proposed by Rigby and Stasinopoulos (2005) and explore whether it has the scope to aide the analysis of income distributions. The structure of this paper is as follows: In the next section, we contrast unconditional income distributions against conditional income distribution, highlighting the need to also analyse the latter. Thereafter, we introduce the parametric modelling approach for the conditional income distribution using the Dagum distribution. In the subsequent section, we line out the methodology of GAMLSS for parameter estimation with a special focus on the estimation of Dagum distributions. In the penultimate section, we estimate the conditional income distributions for males with respect to the three explanatory variables age, educational attainment and region. Using the Kolmogorov-Smirnov test we check whether the Dagum distribution is acceptable for the analysis of our CIDs. We then use three exemplary conditional income distributions to highlight the diverse nature of the evolution of incomes in Germany and show that going beyond conditional mean incomes can provide new insights for the analysis of income inequality in Germany. In the last section we conclude.

2 Unconditional and conditional income distributions

Using the data available in the German Socio-Economic Panel (SOEP) database[2], we consider the labour income as defined in the gross market income definition from Bach et al. (2009). Thereby our income definition entails wage income (including social security contributions) both from the private and the public sector as well business income from agriculture and forestry, unincorporated enterprise and self-employment. However, contrary to Bach et al. (2009) we exclude capital income. Our labour income definition thus entails practically all income derived from the factor labour. Consequently, we implicitly incorporate both changes in wage rates and changes in working time[3]. Since we aim to analyse the evolution of labour related income inequality at large, this seems the most appropriate definition to use. For more elaboration on the data see the appendix.

2.1 Unconditional income distributions

The two income distributions which we will consider in the following are displayed in Figure 1. Note that we inflated incomes in 1992 to be valued in 2010 Euros using the consumer price index (Statistisches Bundesamt, 2012).

[2]For an elaboration of the advantages and shortfalls of the SOEP database see Bach et al. (2009, pp.307-308).

[3]Working time is of high importance as there was a steep rise in unemployment as well as part-time and marginal part-time work in the period under consideration (Biewen and Juhasz, 2012).

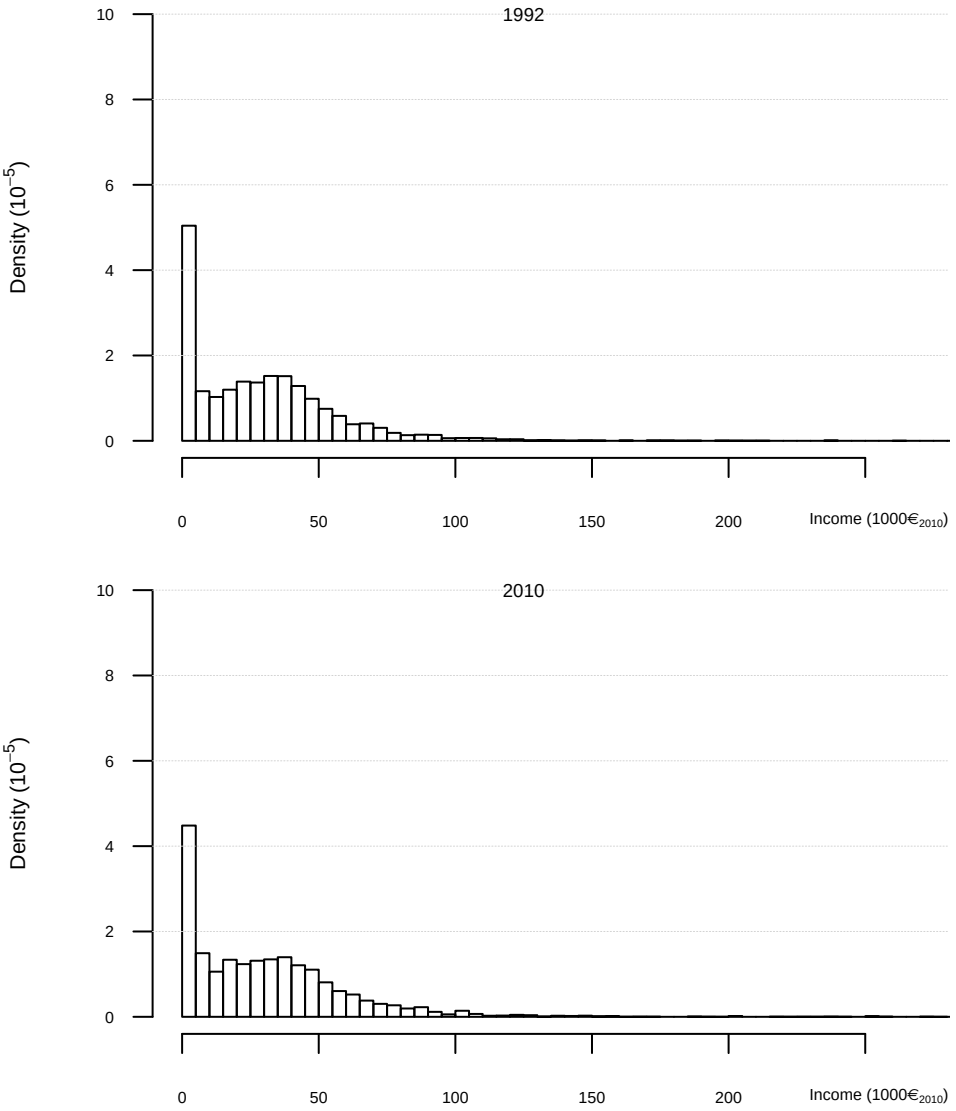

Figure 1: Unconditional German Income Distribution

From the graphic alone, no striking differences can be noted. One can observe a slight fall in the density of the lowest income bin, as the share of people with an income below 5,000€ per annum declines by over ten percent from 1992 to 2010. Concerning the other bins, differences are very slim. Looking carefully one can notice a slight upward shift of the densities from 50,000€ per annum.

Looking at some summary statistics, we observe an increase in the mean income from 29,200€ in 1992 to 31,000€ in 2010. The median also increased in the same time period from 25,400€ to 26,400€. Aggregate inequality, as measured by the Gini coefficient, somewhat surprisingly even falls slightly from 0.503 to 0.497. This implies a 1% fall in inequality as measured by the Gini coefficient and is somewhat counter-intuitive given the literature on rising inequality.

The reason for this slightly puzzling result is found in the concealing nature of our unconditional analysis which amalgamates various developments. Among other things it hides the differences due to developments in East and West Germany, changes in the income distribution due to rising female labour market participation and changes due to the changing demographic structure of the German population.[4] What this highlights is the importance of considering not only unconditional income distributions but to consider conditional income distributions (CIDs), that is to decompose the unconditional income distribution. This is standard in the literature.

Yet standard income decomposition by definition takes a macro-perspective in the sense its primary purpose is the explaining the contribution of various changes to the changes of the aggregate income distribution. The analysis we will conduct in the following differs from classical income inequality decomposition as it takes a fundamentally different micro-perspective. The focus is directed towards the changes of the income distributions of subpopulations. The impact of these changes on the aggregate income distribution is secondary.[5] The main impetus of our research is thus the analysis of changes experienced by subpopulations rather than the population at large.

2.2 Conditional income distributions

Following one of the most popular decomposition categories, namely decomposition by population groups, we will condition our income distributions on various demographic variables - namely region, education and age.[6] We consider region as a binary variable differentiating between the geographical region of the former Federal Republic of Germany and the former German Democratic Republic (entailing both former East and West Berlin).[7] Following Acemoglu (2002) we consider education as a binary variable as well which is unity for everybody who has obtained at least a university degree and zero otherwise. Lastly age is considered in a different manner. In the literature age has generally be considered by a finite number of groups. For example Dustmann et al. (2009) split up their sample into three age groups for their decomposition.[8] Yet, Morduch and Sicular (2002) point out that age should more appropriately be considered as a continuous

[4]See the appendix for a first analysis of conditional income distributions dependent on gender, age and region which highlight the very different dynamics of the development of the conditional income distributions.

[5]GAMLSS based estimation of CIDs can also be used for classical decomposition. Yet this aspect won't be discussed in detail here.

[6]Mainly for reasons of comparability with the existent literature, which unfortunately has focused heavily on the distribution of male incomes/wages, we chose to exclude the conditional income distributions of females. Yet it should be noted that preliminary analyses have shown arguably more interesting dynamics for female incomes than for the male side.

[7]A model with a finer geographical resolution would possibly yield interesting new results and further work should be done on the improved incorporation of geographical information.

[8]For an overview, we provide histograms on the histograms for income distribution conditioned on these three age groups and the other three binary variables in the appendix.

variable.[9] They also point to the problem that as categories or variables are added the number of distributions which need to be estimated increases in a multiplicative manner. Thus given the usual finite number of observations (in the order of thousands or tens of thousands) a direct estimation of the conditional distribution quickly becomes unstable. Consequently regularisation is required. This regularisation is achieved with the GAMLSS approach which is discussed in more detail below. Inherent to the approach which we will pursue is the assumption that we can adequately model the CIDs by a parametric distribution. While we acknowledge that "the use of the parametric approach to distributional analysis runs counter to the general trend towards the pursuit of non-parametric methods, [...]" (Cowell, 2000, .145) we perceive the parametric approach as a form of regularisation itself which by imposing a structure lends stability to the estimation process. Moreover, we concur with Morduch and Sicular (2002, p.93) that it is often "necessary to impose more structure in order to draw sharp conclusion." And last but not least it should be noted that parametric models are better suited for robustness checks (see Silber, 1999, p.8). Naturally, the applicability of any parametric approach hinges on the "agreement between the model being identified and the actual observations" (Dagum, 1977). In other words it is critical to find a parametric model which is able to provide a sufficiently "good fit of the whole range of the distribution" (ibid.) for all the covariate sets of interest. The more diverse the sets of covariates under consideration are, the more the CIDs are likely to differ. With increasingly diverse sets of covariates, more flexible distributions are thus likely to be required.[10]

A lot of ink and paper has been dedicated to the description of the aggregate income distribution of single countries in a parametric manner.[11] Borrowing from this literature we have tried various parametric distributions. One fundamental problem we encountered was the frequently bimodal nature of the CIDs already reported in various other contexts (see Cowell et al., 1996). This structure is due to the inclusion of the whole population in the specified age range irrespective of their employment status.[12] To our knowledge this bimodal structure is not accounted for in any of the standard aggregate income distributions. To take account for this problem we applied a rough-and-ready trisection to our data, truncating all recipients of an income below 4,800€[13] from the

[9] While the difference may become only of academic interest from a certain resolution onwards, the rather coarse differentiation with very few age groups disregards important developments within the age groups. Taking the first age interval of Dustmann et al. (2009) as an example we would have no notion of the direct income distribution effects in the early twenties when vocational training typically ends or the mid/late twenties when students typically leave university. All these important labour market dynamics and the associated changes to the income distribution are ground to analytical dust by the coarse structure of the age categorisation.

[10] Naturally, more flexibility generally requires more parameters which may lead to estimation problems. This aspect is treated in more detail in the GAMLSS section.

[11] Kleiber and Kotz (2003) as well as Chotikapanich (2008) provide an excellent overview.

[12] In this respect our study markedly differs from previous studies with a micro perspective like **?** or Card et al. (2013) who only consider (full-time employed) male workers.

[13] This figure was chosen on grounds of the so called "400€ -jobs" which falls under the category of minor (and consequently atypical) employment which is exempt from social security.

main group, such that the parametric income distribution was only estimated for incomes above this level. The incomes below the level are then in turn divided up into zero and above-zero incomes. Using standard sequential logit estimation we determine the probabilities for a zero income, a low income (i.e. that the income is greater than zero but reaches no more than 4,800€), called precarious income from hereon, or whether the income falls into the income category above 4800€, which we will turn to now.[14]

The natural starting choice for the truncated income distribution was the log-normal distribution, which was popularised by Gibrat (1931). Although its well documented problems to adequately fit the upper bound of the distribution (see among others Atkinson, 1975) should be alleviated by the conditioning on variables like education and age, we found the distribution to be inadequate for our CIDs, too. Also the generalised normal distribution (also known as the Power Exponential distribution) proved a poor fit of the log-incomes under consideration. In general we found that an adequate fit for the log-transformed income to be problematic. The main reason for this is the left-skewed shape of some CIDs. While this skewness can partially be explained by the nature of the log-transformation which in some cases overcompensates the right-skewness of the untransformed conditional income distribution, it is also likely that an underrepresentation of high incomes in the SOEP sample under consideration (see Bach et al., 2009) comes into play. Due to the truncation we described above, the fitting of a parametric left-skewed distribution has proved infeasible. In the following we will therefore solely consider the untransformed incomes as our dependent variable.

Returning to a unity-link we found that other distributions from the exponential family, like the gamma distribution, which are proposed in the literature for aggregate income distributions proved inapt. By contrast distributions of the beta-type which are members of the Pearson system (see Kleiber and Kotz, 2003) proved more promising. With regard to three-parameter distributions we have looked at the Dagum distribution as well as the Burr distribution which is also known as the Singh-Maddala distribution. Using the Kullback-Leibler Divergence as well as the L2 distance measure to a kernel-density estimate[15] of the conditional income distribution we found the former to be more adequate in most cases.[16] of which the former proved more adequate. This finding is echoed for aggregate distributions by Kleiber and Kotz (2003, p.212). As McDonald (1984) points out, the more flexible Generalised Beta distribution of Second Kind with 4 parameters or the even the 5 parameter Generalised Beta distribution might provide an even better fit. However, the additional parameter estimation required for the Generalised Beta distributions has so far proved to be infeasible due to the relatively low number of observations. Thus we will restrain ourselves

[14]The associated mathematical problems with such a rough-and-ready truncation are discussed in Dagum (1977). This emphasises that more elaborate methods are required in the future.

[15]We used an automatic bandwidth selection from Sheather and Jones (1991) and an Epanechnikov kernel.

[16]In general the Dagum distribution provided a better fit in the tails of the distribution while the Burr Distribution was normally better fitting in the centre of the distribution.

to the three parameter Dagum distribution for the moment.

2.3 The Dagum distribution

In the subsequent estimations of the conditional distributions we will therefore primarily consider the Dagum distribution.The distribution was introduced as an income distribution by the Italian Camilo Dagum in 1977. It is nested in the Generalised Beta distribution of Second Kind (Mielke and Johnson, 1974) and its probability density function is given by

$$f(y) = \frac{apx^{ap-1}}{b^{ap}[1+(x/b)^a]^{p+1}},\ y > 0, \tag{1}$$

which yields the cumulative density function

$$F(y) = \left(1 + \left(\frac{x}{b}\right)^{-a}\right)^{-p},\ y > 0, \tag{2}$$

where $a,\ b,\ p > 0$.[17]

Following the notation from Kleiber and Kotz (2003), b is a scale parameter while a and p are shape parameters. As we can see in Equation (1) the parameters impact on the density is intricate as there are strong interrelations among the three parameters. Thus the direct economic interpretation of the parameters is limited. However, as the estimated parameters specify the conditional distribution, we can assess the CID. For this purpose, we able to borrow from the wide range measures applied in the economic literature to assess a size distribution with respect to the question at hand. First and foremost the literature tends to consider the distribution's first moment, i.e. the mean. For the Dagum distribution the first moment is given by:

$$\mu = \frac{bB(p+1/a, 1-1/a)}{B(p,1)}, \quad \text{for } a > 1, \tag{3}$$

where B denotes the beta function as defined in Abramowitz and Stegun (1972, p.258). Note that the moment only exists for $1 < a$. This measure of location generally forms the backbone of econometric analysis and will be considered in detail. Another location measure is the mode, which is given by

$$\text{mode} = b\left(\frac{ap-1}{a+1}\right)^{1/a}, \quad \text{if } ap > 1, \tag{4}$$

and is at zero otherwise. While we will not analyse the mode of conditional distributions in detail,

[17]Note that we are using the notation and parametrisation of Kleiber and Kotz (2003), which is slightly different from the one to the parametrisation of Stasinopoulos and Rigby (2007) whose package we apply for estimation.

it is important to note that the Dagum distribution can thus be both unimodal and zeromodal. The importance of this attribute was already noted by Dagum (1977). We will see later on that for conditional distributions the ability to model zeromodal distributions is a key requirement.

However, while location measures are quintessential to any analysis of income distributions, their account of distributions is naturally limited. Thus additional measures will have to be considered. Thus we consider additional (standardised) moments like the standard deviation and the skewness of the distribution, which are given in Equations 5 and 6 respectively.

$$\sigma = b\sqrt{\frac{B(p+2/a, 1-2/a)}{B(p,1)} - \left(\frac{B(p+1/a, 1-1/a)}{B(p,1)}\right)^2}, \quad \text{for } a > 2, \tag{5}$$

$$\text{skewness} = \frac{B^2(p,1)\lambda_i - 3B(p,1)\lambda_2\lambda_1 + 2\lambda_1^3}{[B(p,1)\lambda_2 - \lambda_1^2]^{3/2}]}, \quad \text{for } a > 3, \tag{6}$$

where $\lambda_i = B(p + i/a, 1 - i/a)$. For $a \leq 2$ and $a \leq 3$ the second and third moment do not exist respectively. In addition to these moments we also consider standard inequality measures which comprise both these aspects.

The two most widely used inequality measures are the Gini coefficient and the the of generalised entropy measures. The Gini coefficient can easily be obtained by

$$\mathrm{G} = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1, \tag{7}$$

where Γ denotes the gamma function as defined by Abramowitz and Stegun (1972, p.255).

Similarly we can obtain the generalised entropy measures. In the following we will concentrate on the Theil index which is given by

$$\mathrm{I}(1) = \frac{\psi(p+1/a)}{a} - \frac{\psi(1-1/a)}{a} - \Gamma(p+1/a) - \Gamma(1-1/a) + \Gamma(p) + 1, \tag{8}$$

where $\psi(z) = \frac{d}{dz}\log\Gamma(z)$ is the digamma function (see Jenkins, 2007).

Interquantile ranges are the third inequality measure we will consider, which can be obtained directly from Equation (2). In specific, we follow Blau and Kahn (1996) and consider the differential between the 90^{th} and the 50^{th} as well as between the 50^{th} and 10^{th} percentile.

Next to these inequality measures, which by definition are relative concepts, we will also an "absolute" income distribution measure. Thereby we use the c.d.f. to determine the shares of the

subpopulation with an income greater than the following thresholds: 0, 19,350€[18] and 46,000€[19].

Before we turn to the "daunting estimation challenge"(Fortin et al., 2011, p.56) of determining whole conditional distribution we would like to re-emphasize the merit of doing so. Describing any income distribution solely by a one-dimensional measure of descriptive measure of location, dispersion, inequality, etc. implies a loss of information by its very nature. An analysis which evaluates CIDs by these single measures is thus prone to neglect or even negate important aspects of the income distribution. Thus the ability to specify CIDs in their complex distributional nature constitutes an important improvement to the analysis of incomes in general and the analysis of income inequality in specific.

3 GAMLSS estimation of the Dagum distribution

Generalised Models of Location, Scale and Shape (GAMLSS) first proposed by Rigby and Stasinopoulos (2005) provide a class of regression models allowing to adequately model CIDs. Let the CID be given by a parametric density

$$f(y \mid \theta_1, \ldots, \theta_K). \tag{9}$$

Then each of the K parameters θ_k is expressed as an additive composition of the explanatory variables. More specifically we write:

$$g_k(\theta_k) = \eta_k = X_k\beta_k + \sum_{j=1}^{J_k} s_j(z_{j,k}), \tag{10}$$

where g_k is a monotonic link function, X_k is a design matrix containing the explanatory variables considered as linear effects and $z_{j,k}$ denotes the j-th covariate considered as non-linear effects. β_k notifies the corresponding estimator for x_k and $s_j(z_{j,k})$ is the smooth function of the $j-th$ continuous covariate considered in a non-linear way.

[18]This is the annual gross market income as would be obtained if a German full-time employee (35 hours) would be paid at the level of the French minimum wage (salaire minimum interprofessionel de croissance) for the year 2010.

[19]This is the amount of money which Keynes ascribes to be enough to turn the human mind away from pecuniary requirements (see Skidelsky, 2010, p.142).

Applied to our purpose we select the following set-up using the Dagum distribution as CID:

$$\log(a) = \eta_a = s_{1a}(age) + Hs_{2a}(age) + Es_{3a}(age) + HEs_{4a}(age), \tag{11}$$

$$\log(b) = \eta_b = s_{1b}(age) + Hs_{2b}(age) + Es_{3b}(age) + HEs_{4b}(age), \tag{12}$$

$$\log(p) = \eta_p = s_{1p}(age) + Hs_{2p}(age) + Es_{3p}(age) + HEs_{4p}(age), \tag{13}$$

where s denotes a smooth function such that we model the effect of age in a non-linear way. We thereby follow the notion of Lemieux (2003) that the relation between earnings and experience (or age) is not linear. H is a binary variable which is unity if we consider the CID for people with higher education. E is also a binary variable which is unity if the CID is for people living in the Eastern part of Germany. The log link thereby implies that additive structure of the explanatory variables is transformed to a multiplicative structure.[20]

The frequentist estimation procedure implemented in the `gamlss` package in `R` and is described in detail by Rigby and Stasinopoulos (2005). Naturally, as these models are highly complex, the are several pitfalls which are considered in Section C in the appendix. Bearing those aspects in mind, we turn to the estimation of the CIDs for males in the years 1992 and 2010.

4 Zooming in on intra-group income inequality in Germany

As pointed out in Section 2.2, we model income distributions in a two-step procedure, such that point masses for zero-incomes and the truncated low incomes are truncated from the rest of the income distribution which is considered by GAMLSS. This implies that we have five parameters for each whole CID (2 for the point masses and three for the Dagum distribution which is used to model the truncated CID). In this section we display the estimates we obtained and use both parameters and auxiliary measures for the resultant CID to analyse inter-group as well as inter-temporal differences of the CIDs.

These studies regard CIDs of males for the years 1992 and 2010. For each estimation result we will proceed in the following manner: First we will display the parameter estimates as obtained in a frequentist framework. Using these parameter estimates, we will use the Kolmogorov-Smirnov test to check whether our Dagum-based parametric model is sufficiently close to the empirical reality. The test results are displayed in Section E.2 in the appendix and contrasted against results using

[20] Lemieux (2003) notes that for the standard Mincer-wage equations, which we can assume to underlie our CIDs, a multiplicative model set-up would be appropriate. However, due to our specific form of parametrisation in this paper this property of the log-link is irrelevant.

the log-normal distribution. Subsequently, we go on to do some exemplary analysis of the CIDs. However, the main purpose of our current inquiry is not inference with respect to the size and direction of the effect of specific explanatory variables but rather to show that GAMLSS can be used to model CIDs- Hence, we will refrain from a comprehensive analysis of the effect of age, education, region and gender on the CID. For each period and gender we select three subpopulations with respect to the explanatory variables and use them show how the conditional income distributions obtained by expand the scope for analysis. The subsequent analysis should thus first and foremost be understood as a descriptive analysis with the scope of also conducting inference in the future. For illustrative purposes we will consider the following three subpopulations:

- 25-year old men living in the Eastern part of Germany without higher education - abbreviated by subpopulation M1 (SM1) from here on.
- 37-year old men living in the West without higher education - subpopulation M2 (SM2) from here on.
- 50-year old men living in the West with higher education - subpopulation M3 (SM3) from here on.

We do so following our three stage estimation strategy. Firstly we discuss the conditional probability for the zero and precarious incomes on the one hand and the truncated distribution of higher incomes on the other hand. Subsequently we proceed to discuss the latter in some more detail before we also include the former and analyse the whole CIDs.

4.1 Macroeconomic background for income distributions in 1992

In 1992 the economic shock waves of the integration of Eastern Germany were still reverberating. While in Western Germany the economy was starting to cool down, the expected "strong expansion" of output (and employment) did not materialise in 1992(see Sachverständigenrat, 1992, pp.71-124) with real GDP growth for the whole of Germany around 2% (Statistisches Bundesamt, 2014b). Nonetheless employment in the West was still stable with the unemployment throughout 1992 below 7%. By contrast in the East unemployment had risen sharply from 10% in 1991 to nearly 15% in 1992 (Statistisches Bundesamt, 2014a). Also short time work and other measures depressed incomes from labour. With this macroeconomic backdrop in mind we turn to the analysis of dependent income distributions.

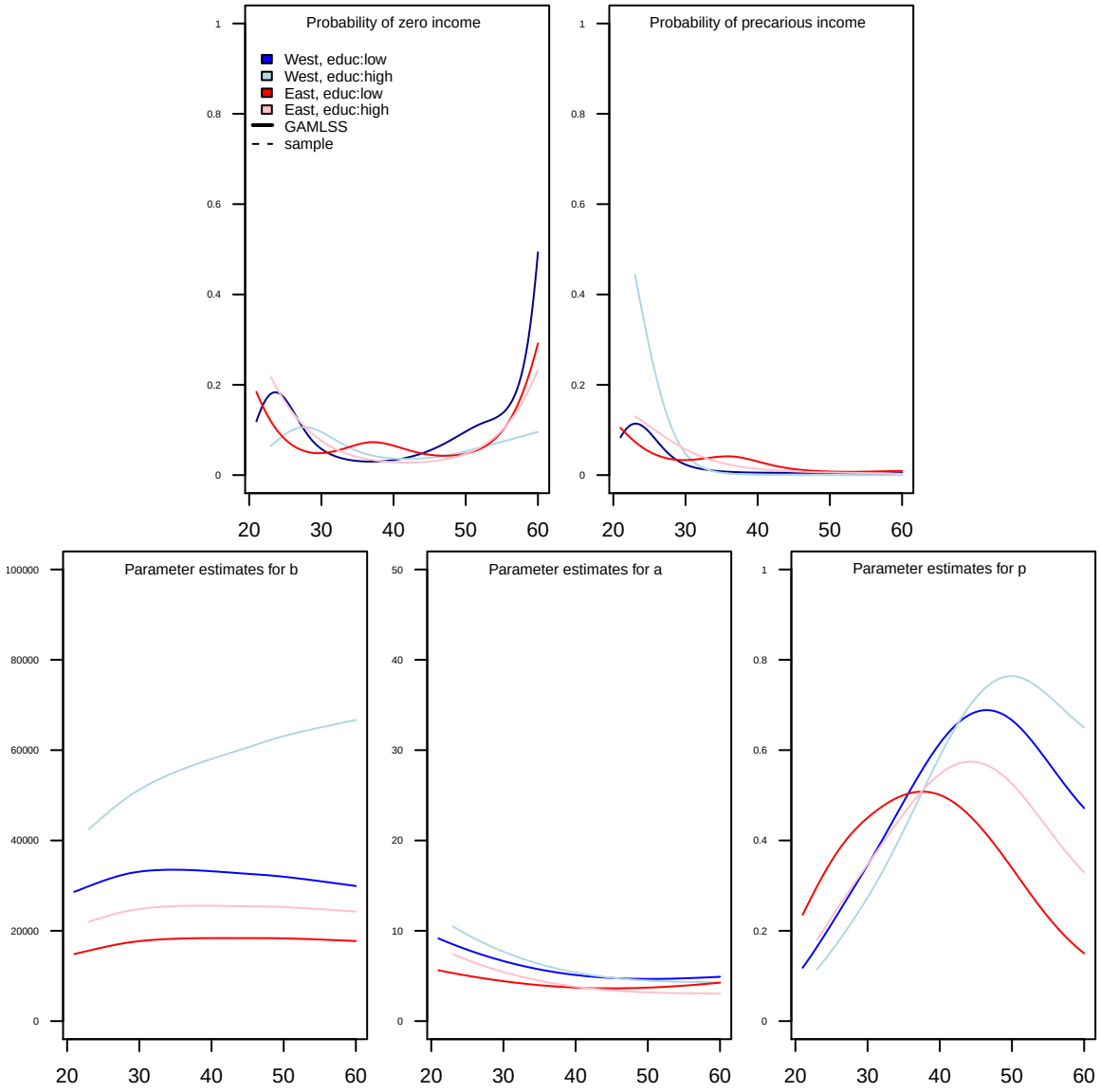

Figure 2: Parameter estimates for Males in 1992

4.2 Conditional income distributions of males in 1992

Figure 2 displays the parameters estimates.[21] From the parameter estimates for zero-incomes we can observe the expected pattern of an inverted U-shape, as no-employment situations are more frequent at a young age during or directly after education as well as towards the end of the age span as retirement sets in. For SM1 we observe the highest share of zero incomes (8%) compared to the level of SM2 (3%) and SM3 (5%). Similarly, for the share of precarious incomes we have also have a higher share for SM1 (5%) than for SM2 (1%) and SM3 (0%). Combining these two

[21]For higher education we only display the parameters from the age of 23 onwards as beforehand very few students are likely to have completed their higher education degree.

measures our estimates thus portray a picture whereby the probability mass below 4,800€ is much higher in the CID SM1 than in those of SM2 and SM3. As pointed out before, the direct interpretation of the parameters of the Dagum distribution is difficult. Yet what can be observed is that the scale parameter b is greatest for SM3 and smallest for SM1, indicating greater mean and standard deviation c.p., i.e. for given parameter values a and p. For a more comprehensive analysis of the truncated CID by auxiliary measures obtained from the parameter estimates see Section E.1 in the appendix.

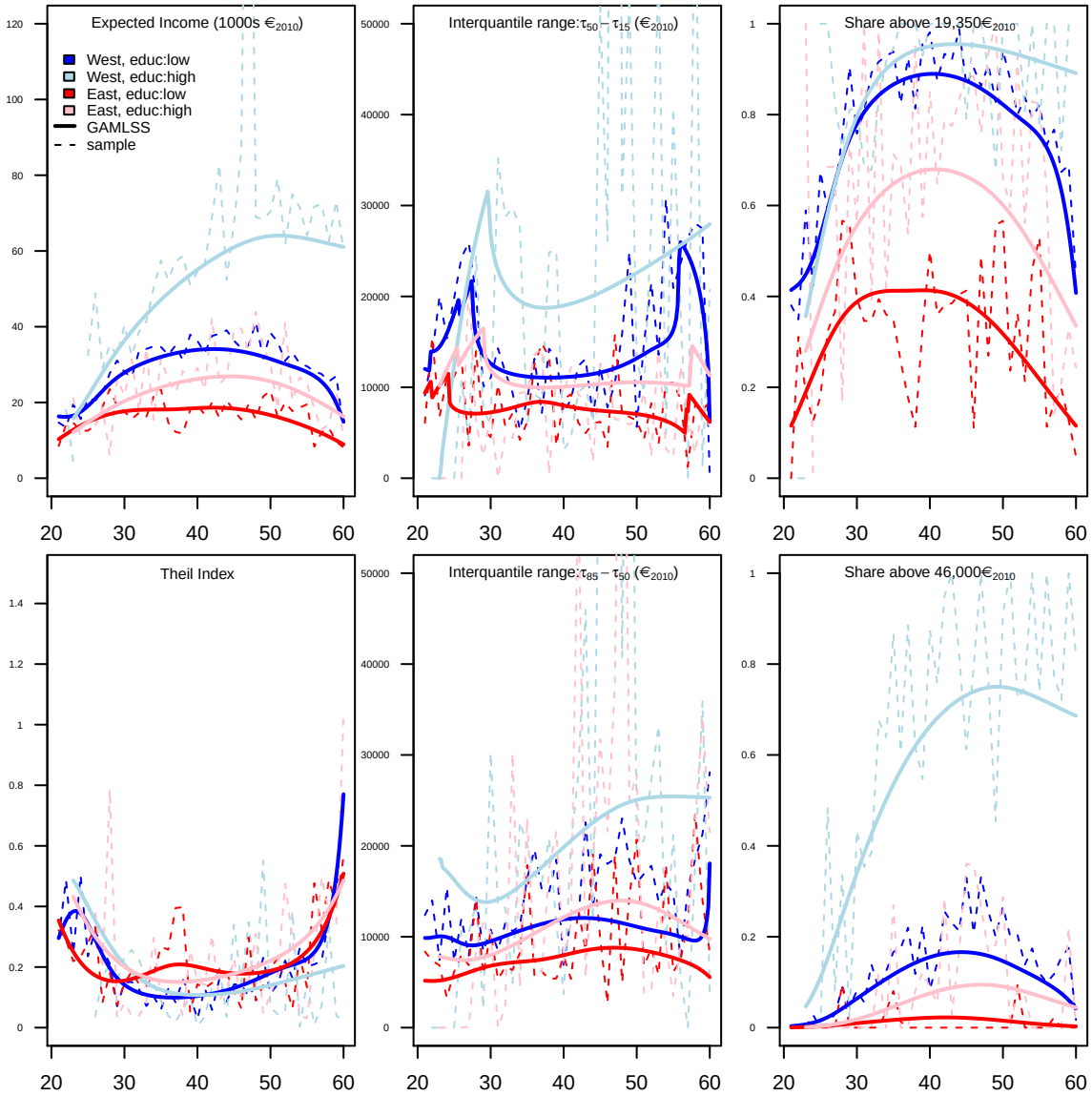

Figure 3: Measures for whole CIDs of males in 1992

Figure 3 displays six measures on the whole conditional distributions we estimated. The thick lines display the resulting measure from the GAMLSS estimation of the conditional distribution, while the dashed lines display the measure obtained directly from the sample for a given age, education

level and region. On the top left we display the conditional expected incomes. The general results are little surprising showing the positive relation between education and income, age (a proxy for experience) and income as well as higher incomes in the West. For our subpopulations we observe the highest mean income for SM3 (64,000€) and the lowest for SM1 (14,800€). This analysis is standard. More interesting for our analysis is the other attributes of the CID.

In order to get a grasp of the within-group inequality we use the Theil Index as well as two interquantile ranges. As can be observed we generally have a U-shape which is little surprising given the stark differences due to large parts of the subpopulations in education and retirement for the two extremes of the age-range respectively. For SM1 we have a low within-group inequality (0.19) compared men of the same age but different regions and with education levels we observe higher within-group inequality as measured by the Theil Index. Using the interquantile ranges this finding is echoed.[22] Yet compared with SM2 (Theil index= 0.10), SM1 portrays a higher Theil Index while the interquantile ranges are higher for SM2. For SM3 we get a slightly higher value for the Theil index (0.14) but much greater values for the interquantile ranges. The differences in these inequality measures can in part be explained by the missing scaling for the interquantile ranges. Yet closer observation also highlights the multi-layered nature the differences in the inequality measures for the three CIDs (see Section E.1). For the truncated CID, we observe that inequality as measured by the coefficient of variation for SM3 (0.45) is only slightly higher than for SM1 (0.40) and it also portrays a higher skewness (2.9 and 1.6 respectively). Consequently, the inequality in the truncated CID, as measured by the Theil index, is slightly higher for SM3 (0.08) than for SM1 (0.07). And it is lowest for SM2 (0.06). Yet, with the inclusion of the zero- and precarious incomes the scales are tilted to attribute the highest Theil index to SM1. For a further analysis of the impact of this inclusion for the whole CID with regard to some other distribution measures, see the appendix (not yet written down). This highlights that inequality, as measured by the Theil index, can be very different for a given subpopulation depending on whether we consider just the 'core' labour market (where zero- and precarious incomes are excluded) or the incomes of every member of that subpopulation. Hence our analysis shows that with regard to the 'core' labour markets for the various subpopulations shows to be least inequitable for SM2. Given the fact that union strength is presumably greatest among SM2 this is hardly surprising. Yet considering the differences between SM1 and SM3 our analysis highlights the discrepancy between inequality within the labour market and overall income inequality for the subpopulations.

For the 'absolute' inequality measures we can observe the share of that is below the 19,350€ is highest for SM1 and lowest for SM3, which is little surprising given the mean income levels well documented in the literature. Also the share of people above the 46,000€ is as expected highest

[22]The spikes in the graphic are due to our parametrisation of precarious incomes as a point mass, such that below 4,800€ the income distribution is not continuous.

for SM3 and lowest for SM1. Indeed for SM1 this threshold is practically out of reach in 1992.

4.3 Macroeconomic background for income distributions in 2010

In 2010 the recovery after the economic crisis in the previous years onwards was fully under way. Contrary to most other OECD countries the German labour market proved to be relatively robust during the crisis and such that unemployment did not increase dramatically over the course of the crisis (Sachverständigenrat, 2010). In 2010 the unemployment rate stood at 8.6% (Statistisches Bundesamt, 2014a) while real GDP bounced back from the 5.1% decline in the previous year, with a growth of 4% in 2010 (Statistisches Bundesamt, 2014b).

4.4 Conditional income distributions of males in 2010

Contrary to the previous analysis we will now first and foremost concentrate on the inter-temporal differences rather than the intra-temporal differences.

Looking at the parameters we can observe that the share of people with zero-income has increased dramatically for SM1 (19%), while it more or less remains unchanged for SM2 (4%) and has fallen slightly for SM3 (3%). With regard to the share with a precarious income, we observe a rise of similar magnitude for SM1 (16%) while for SM2 (3%) and SM3 (0%) things have also developed similarly as for zero-incomes. Combining these two measures, we can thus see a drastic rise in the share of people with an income below 4,800€ for SM1, while for SM2 and SM3 things have not deteriorated in the same manner. For the parameters of the Dagum distribution we gain refer to Section E.1 in the appendix and go on to the auxiliary measures for the whole CID.

The most noticeable difference of the expectation for the dependent income distributions displayed on the top left of Figure 5 is the surpassing of the average incomes of men with higher education in the East of those without higher education in the West. Also the rift of incomes between East and West for men without higher education has also narrowed. This convergence in mean incomes already pointed out by Vollmer et al. (2013) and others. Especially for the older generations in the East the increase in income dispersion is dramatic. For those with higher education this is combined with a persistent high skewness again causing this group to show high within-inequality as indicated by the Theil index. With respect to the convergence of the whole income distribution (rather than just the mean) one has to concede that true convergence is still some way off. For the truncated income distribution, mean convergence across the age groups is driven by the catching up of incomes for some with a large part of the male population in the East left wanting as indicated by the higher standard deviation and skewness.

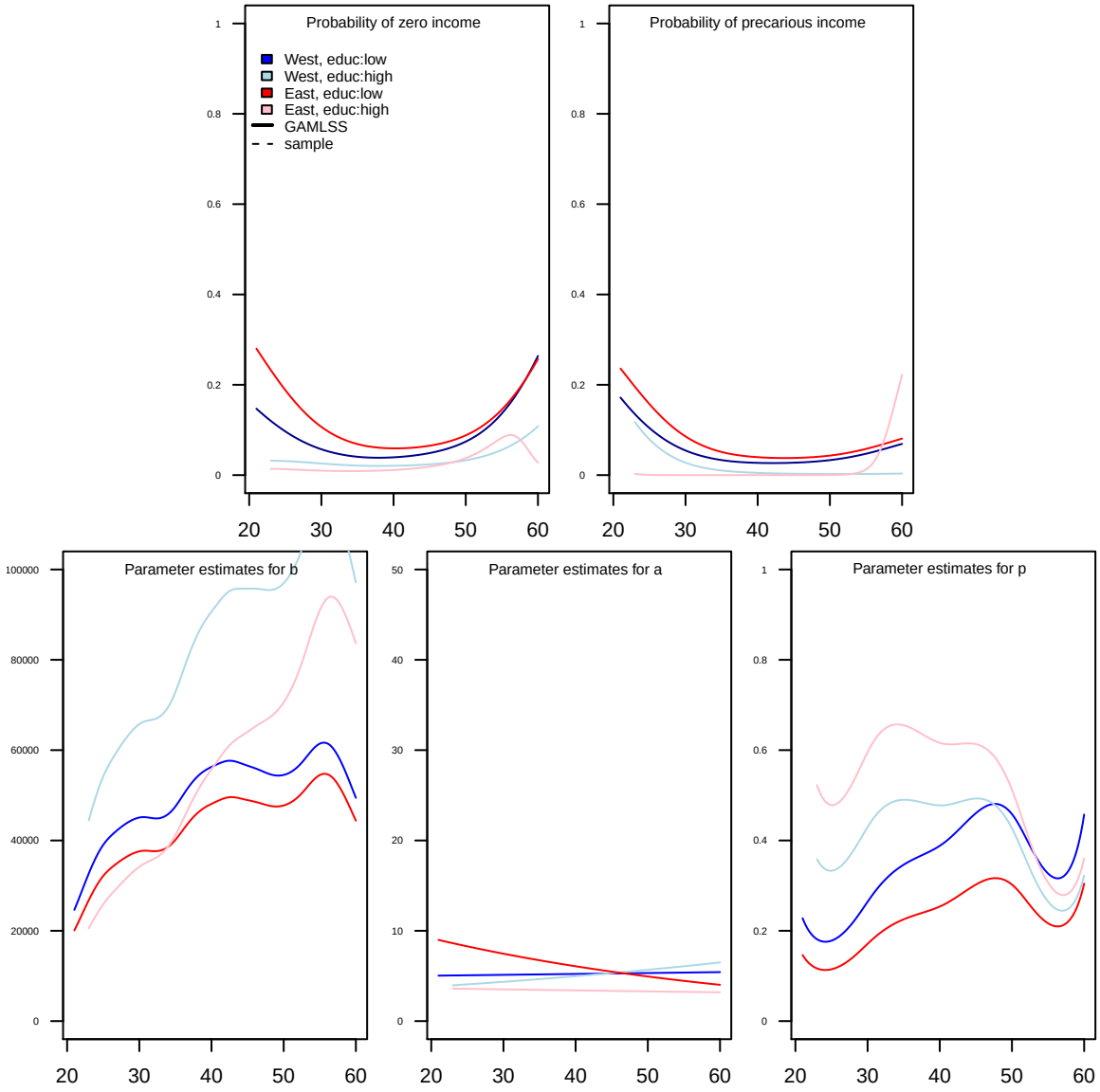

Figure 4: Parameter estimates for Males in 2010

The most noticeable difference of the expectation for the dependent income distributions displayed on the top left of Figure 5 is the surpassing of the average incomes of men with higher education in the East of those without higher education in the West. Also the rift of incomes between East and West for men without higher education has also narrowed. This convergence in mean incomes already pointed out by Vollmer et al. (2013) and others. For SM2 and SM3 we find the expected income to have risen to 41,300€ and 80,700€ respectively. By contrast for SM1 we find a slight decline to 14,100€. This decline of the mean of the CID of SM1 is mostly driven by the dramatic increase of zero- and precarious incomes discussed above. For the truncated CID we observe an increase of the mean from 12,200 to 16,400 as well as a rise of the Theil index from 0.07 to 0.13. Combined these changes result in a rise of inequality as measured by the Theil index for SM1 (0.48)

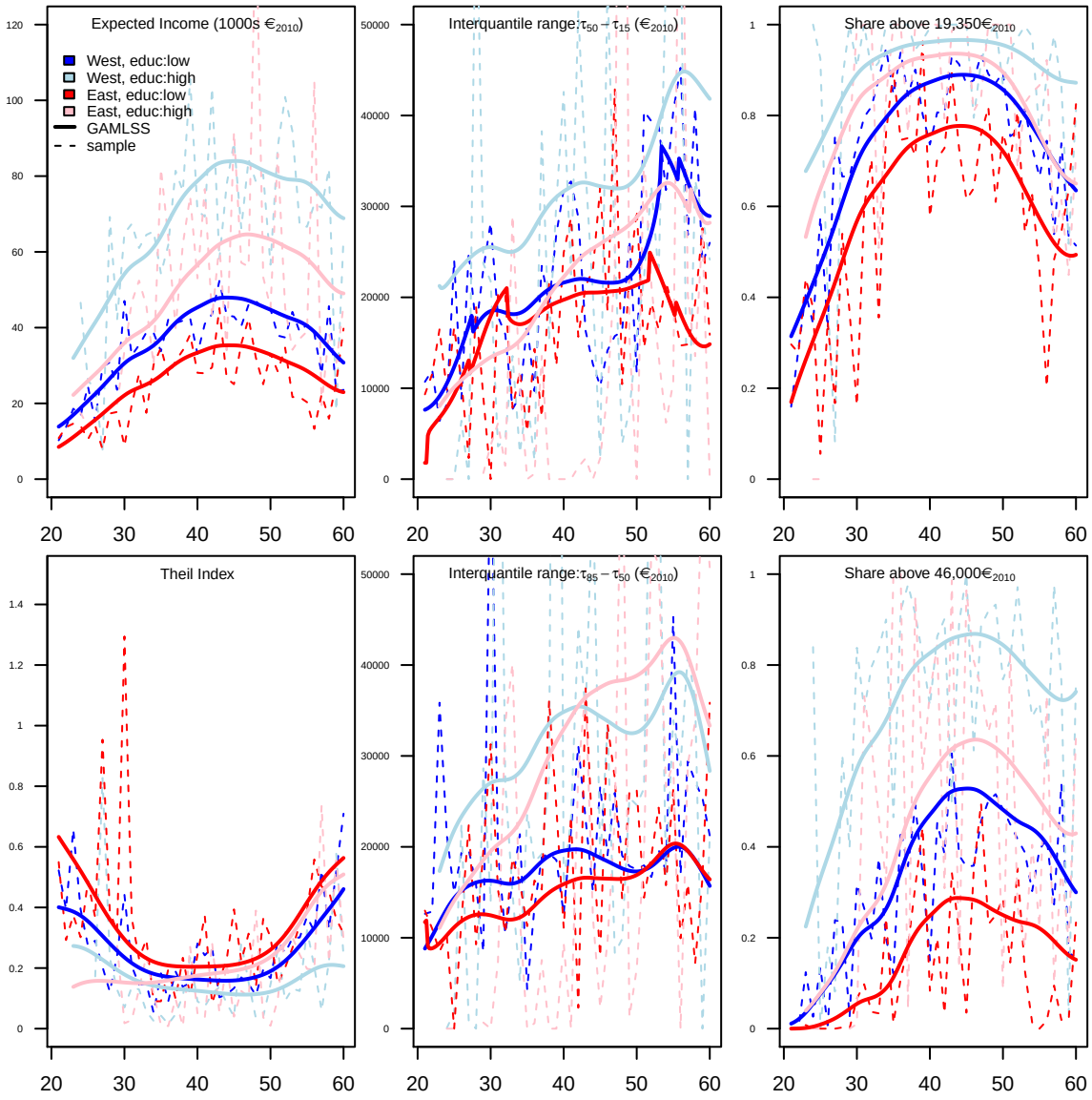

Figure 5: Measures for whole CIDs of males in 2010

which is an expression of a diverging income structure within the subpopulation where larger parts are trapped in the zero- or precarious income region while the incomes above 4,800€ while also experiencing growing dispersion have risen considerably on average. For SM2 and SM3 the rising mean incomes accompanied by an increased and decreased Theil index respectively. As we already note is that there is no substantial inflation of zero- and precarious incomes as for SM1. The main impetus driving inequality differences is thus to be found in the truncated CID. For SM2 we observe a rise in the Theil index to 0.17 which is mostly driven by the rising dispersion in the centre of the truncated CID, as can be seen by the sharply rising interquantile ranges or by the coefficient of variation. By contrast for SM3 we observe a falling coefficient of variation as well as a reduced skewness within the truncated CID. In other words, for SM3 we see a location shift, without

increasing dispersion as well as reduced weight to extremely high incomes, reducing the Theil index. Concerning the thresholds, the development is generally positive as greater percentages are above both thresholds for most subpopulations. For SM1 we observe a slight increase for both thresholds (26% to 34% and 0% to 1%). For SM2 the growing dispersion has lead to a reduced share for the first threshold (88% to 85%) and an increase for the second threshold (13% to 39%). For SM3 the high level above the first threshold was held constant at 95% with a substantial increase for the share above the second threshold (75% to 84%). Our results thus show that the evolution of incomes of males in Germany has been very diverse in its nature as CIDs developed very differently. While much more additional research on this matter is needed, this finding hints at the possibility that there exists an additional aspect of skill-biased technological change. Not only do skill-premiums (the difference of conditional mean incomes of skilled) seem to have grown, also the shape of the CIDs seems to change. While the income distributions of young lowly skilled workers have tended to become more skewed with large proportions out of employment or only on very low incomes and ever fewer members of this strata gaining access to the well-paid jobs at disposal, the magnitude of skill-bias is gravely underestimated by a mean centred portrayal. By contrast, in the higher better educated strata, a development of contracting CID is indicated such that not only the centres of the CIDs, i.e. the means, seem to move apart but also the probability masses within those distributions seem to move in opposite direction so that ever thinner tails of the distributions are overlapping.

5 Conclusion

At the outset of this article we highlighted the need for the analysis of conditional income distributions, not only as building blocks for the aggregate income distribution of a nation but as objects of analysis in their own right. Yet even with considerably large databases like the SOEP we are quickly running into severe estimation difficulties as we attempt condition on a set of explanatory variables. The semiparametric additive structure provided by GAMLSS regression techniques allow for regularisation and consequently aide the modelling of conditional income distributions. For the two periods under consideration we found that conditional income distributions of males could generally be modelled using the three parameter Dagum distribution. Using the estimates we compared the conditional income distributions of three subpopulations for 1992 and 2010. We found that while 25-year old males without higher education in the East in regular employment experienced rising mean incomes, the shape of the income distribution changed dramatically as lack of employment and precarious employment are much more severe in 2010. For 37-year old males without higher education in the West we find that while zero-incomes and precarious incomes have been kept in check, income dispersion has increased dramatically leading to a substantial rise in

within-group inequality. Conversely, for 50-year old males with higher education, not only mean incomes have increased substantially but also a decreasing dispersion among their incomes can be observed. While further analysis in this direction is needed, such an analysis of conditional income distributions hint at additional possible aspects aggravating the extent of the much discussed skill-biased technological change.

In the past two decades the analysis of income distributions and income inequality has come back from the cold fringes of the economic literature to the centre of scientific debate according to Atkinson (1997). It is our belief that for a comprehensive analysis of income inequality, we must consider the full nature of the change of the income distribution(s) with respect to a set of variables. The aim of this paper was to assess the principle capability of GAMLSS to model CIDs. The results we found are promising, indicating that with the correct distributional choice GAMLSS have the scope to do it. But much more work needs to be done to see whether GAMLSS provide the blanket coverage which is required to not only get the analysis of income inequality back in from the cold but to get it warmly tucked into the right modelling framework.

References

M. Abramowitz and I. A. Stegun (1972): Handbook of Mathematical Functions, Dover, New York.

D. Acemoglu (2002): Technical Change, Inequality and the Labor Market, in: Journal of Economic Literature, 40(1), pp. 7–72.

H. Akaike (1983): Information measures and model selection, in: Bulletin of the International Statistical Institute, 50, pp. 277–290.

A. B. Atkinson (1975): The economics of inequality, Clarendon Press, Oxford.

— (1997): Bringing Income Distribution in from the Cold, in: Economic Journal, 107(March), pp. 297–321.

— (2003): Income Inequality in OECD: Data and Explanations, CESifo Working Paper 881.

S. Bach, G. Corneo and V. Steiner (2009): From Bottom to Top: the entire Income Distribution in Germany, 1992-2003, in: Review of Income and Wealth, 55(2), pp. 303–330.

BfAS (2013): Der Vierte Armuts- und Reichtumsbericht der Bundesregierung, Bundesministerium für Arbeit und Soziales, Berlin.

P. J. Bickel and M. Rosenblatt (1973): On some Global Measures of the Deviations of Density Function Estimates, in: The Annals of Statistics, 1(6), pp. 1071–1095.

M. Biewen and A. Juhasz (2012): Understanding Rising Income Inequality in Germany, in: Review of Income and Wealth, 58(4), pp. 622–647.

F. D. Blau and L. M. Kahn (1996): International Differences in Male Wage Inequality: Institutions versus Market Forces, in: Journal of Political Economy, 104(4), pp. 791–836.

K. Brachmann, A. Stich and M. Trede (1996): Evaluating Parametric Income Distribution Models, in: Allgemeines Statistisches Archiv, 80(3), pp. 285–298.

D. E. Card, J. Heining and P. Kline (2013): Workplace Heterogeneity and the Rise of German Wage Inequality, in: Quarterly Journal of Economics, 128(3), pp. 967–1015.

V. Chernozhukov, I. Fernandez-Val and B. Melly (2013): Inference on Counterfactual Distributions, in: arXiv:, 0904.0951v6[stat.ME].

D. Chotikapanich (2008): Introduction, in: D. Chotikapanich (ed.), Modeling income distributions and Lorenz curves, pp. ix–xii, Springer, New York.

F. Cowell (2000): Measurement of Inequality, in: A. B. Atkinson and F. Bourguignon (eds.), Handbook of income distribution, pp. 87–166, Elsevier, Amsterdam.

F. A. Cowell, S. P. Jenkins and J. A. Litchfield (1996): The Changing Shape of the UK Income Distribution: Kernel Density Estimates, in: J. Hills (ed.), New inequalities, pp. 49–75, Cambridge University Press, Cambridge.

C. Dagum (1977): A New Model of Personal Income Distribution: Specification and Estimation, in: Economie Applicée, 30, pp. 413–437.

C. Domański and A. Jedrzejczak (1998): Maximum likelihood estimation of the Dagum model parameters, in: International Advances in Economic Research, 4, pp. 243–252.

C. Dustmann, J. Ludsteck and U. Schönberg (2009): Revisiting the German Wage Structure, in: Quarterly Journal of Economics, 124(2), pp. 843–881.

P. H. C. Eilers and B. D. Marx (1996): Flexible Smoothing with B-splines and Penalties, in: Statistical Science, 11(2), pp. 89–102.

B. Fitzenberger (1999): Wages and employment across skill groups: An analysis for West Germany, Physica-Verlag, Heidelberg.

N. M. Fortin and T. Lemieux (1998): Rank regression, wage distributions and the gender gap, in: Journal of Human Resources, 33(3), pp. 610–643.

N. M. Fortin, T. Lemieux and S. Firpo (2011): Decomposition Methods in Economics, in: O. Ashenfelter and D. E. Card (eds.), Handbook of Labor Economics, Vol. 4A, pp. 1–102, North-Holland, Amsterdam.

N. Fuchs-Schündeln, D. Krueger and M. Sommer (2010): Inequality trends for Germany in the last two decades: A tale of two countries, in: Review of Economic Dynamics, 13(1), pp. 103–132.

R. Gibrat (1931): Les Inégalités Economiques, Sirely, Paris.

M. M. Grabka (2013): Codebook for the PEQUIV File 1984-2011, DIW Berlin, Berlin.

J. Hartung, B. Elpelt and K.-H. Klösener (1987): Statistik: Lehr- und Hanbuch der angewandten Statistik, 6 ed.

S. P. Jenkins (2007): Inequality and the GB2 income distribution, in: IZA Discussion paper series, No. 2831.

D. N. Joanes and C. A. Gill (1998): Comparing Measures of Sample Skewness and Kurtosis, in: The Statistician, 47(1), pp. 183–189.

C. Kleiber and S. Kotz (2003): Statistical size distributions in economics and actuarial sciences, Wiley, Hoboken.

N. Klein, M. Denuit, S. Lang and T. Kneib (2014): Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape, in: Insurance: Mathematics and Economics, 55(March), pp. 225–249.

R. Koenker and G. Bassett (1978): Regression quantiles, in: Econometrica, 46(1), pp. 33–50.

P. R. Krugman (2007): The conscience of a liberal, W.W. Norton & Co., New York, 1 ed.

T. Lemieux (2003): The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings, in: Center for Labor Economics -University of California, Working Paper No. 62.

J. Machado and J. Mata (2005): Counterfactual Decompostion of Changes in Wage Distributions Using Quantile Regression, in: Journal in Applied Econometrics, 20(4), pp. 445–465.

J. B. McDonald (1984): Some Generalized Functions for the Size Distribution of Income, in: Econometrica, 52, pp. 647–663.

P. W. Mielke and E. S. Johnson (1974): Some generalized distributions of the second kind having desirable application features in hydrology and meteorology, in: Water Resources Research, 10, pp. 223–226.

J. Morduch and T. Sicular (2002): Rethinking Inequality Decomposition, with Evidence from Rural China, in: The Economic Journal, 112(476), pp. 93–106.

R. A. Rigby and D. M. Stasinopoulos (2005): Generalized Additive Models for Location, Scale and Shape, in: Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), pp. 507–554.

Sachverständigenrat (1992): Jahresgutachten 1992/93 des Sachverständigenrates zur Begutachtung der gesamtwirtschaftlichen Entwicklung, in: Drucksache, 12(3774).

— (2010): Chancen für einen stabilen Aufschwung, Statistisches Bundesamt, Wiesbaden.

S. J. Sheather and M. C. Jones (1991): A reliable data-based bandwidth selection for kernel density estimation, in: Journal of the Royal Statistical Society: Series B (Statistical Methodology), 53(3), pp. 683–690.

J. Silber (ed.) (1999): Handbook of income inequality measurement, Kluwer Academic, Boston.

R. Skidelsky (2010): Keynes: The return of the master, PublicAffairs, New York, 1 ed.

D. M. Stasinopoulos and R. A. Rigby (2007): Generalized Additive Models for Location, Scale and Shape (GAMLSS) in R, in: Journal of Statistical Software, 23(7), pp. 1–46.

Statistisches Bundesamt (2012): Periodensterbetafeln für Deutschland: Allgemeine Sterbetafeln, abgekürzte Sterbetafeln und Sterbetafeln, Statistisches Bundesamt, Wiesbaden.

— (2014a): Bruttoinlandsprodukt, Bruttonationaleinkommen, Volkseinkommen - Lange Reihen ab 1950, in: Volkswirtschaftliche Gesamtrechung, Januar 2014.

— (2014b): Registrierte Arbeitslose, Arbeitslosenquote nach Gebietsstand, URL https://www.destatis.de/DE/ZahlenFakten/Indikatoren/LangeReihen/Arbeitsmarkt/lrarb003.html.

M.-P. Victoria-Feser (1995): Robust methods for the analysis of income distribution models with applications to Dagum's model, in: C. Dagum and A. Lemmi (eds.), Income distribution, welfare, inequality and poverty, pp. 225–239, JAI Press, Greenwich.

S. Vollmer, H. Holzmann, F. Ketterer and S. Klasen (2013): Distribution dynamics of regional GDP per employee in unified Germany, in: Empirical Economics, 44(2), pp. 491–509.

A Data

A.1 Sample

For our analysis we use the SOEP Database. In order to avoid distortion, we follow Bach et al. (2009) and only employ samples A-F, both for 1992 and 2011. We thus explicitly exclude the high-income sample, which allows us account for the upper tail of the distribution more accurately.

It also should be noted that our cross-sectional approach has several weaknesses. As Atkinson (2003) points out, single years are can be bad representatives of longer periods such as decades and can be highly misleading. Secondly, we our cross-sectional approach does not allow us to exploit the panel structure provided by the SOEP. In the future we plan to incorporate fixed and random effects in our analysis of the German income structure. But since many of the recent studies on the German income distribution also use cross-section, we consider a cross-sectional approach to be worthwhile.

Bach et al. (2009) only consider adults of 20 years or older. We extend this age restriction such that we only consider adults between 21 and 60 years of age. The reasoning behind this restriction is that we want to observe changes in the annual gross income over the period of standard employment. We thus exclude the time when most people finance themselves largely by pension payments, very similar to the exclusion of young people who are financed by their parents.

A.2 Gross Market Income

We employ the definition proposed by Bach et al. (2009), excluding capital incomes though. Thereby, only the first two of the following three income components are incorporated to add up the individual's income.

- *Wage income* is the payment of wages and salaries received by the individual from all his employers as well as the employers' social security contributions. To obtain the annual income from wages and salaries we use the Variable *I11110$$* from the PEQUIV File (Grabka, 2013), which entails all the income from a dependent employment relationship.[23] For employees in the private sector we then add 20% in 1992 and 2011 to account for the employers' social security contributions.For civil servants we account for their *Vollversorgung* by adding 47% to their annual labour income for both 1992 and 2011. It should be noted that these are momentary rough-and-ready measures, which we intend to refine later.

[23]Note that we took the income information for a given year from the subsequent SOEP questionnaire, as the annual income in a questionnaire is naturally given for the previous year.

- *Income from business activity* entails includes income from unincorporated business enterprise and from self-employment activities, as well as taxable income from agriculture and forestry. This source of income is also accounted for in *I11110$$* from the PEQUIV File, or rather *ISELF$$* therein.
- *Capital income* entails incomes from interest and dividends as well as renting and leasing and is not considered.

B Graphical analysis of dependent income distributions

In this section we display the graphics, of the decomposition following a two level education definition from Acemoglu (2002) and the three age groups from Dustmann et al. (2009) for 1992 and 2010 for East and West. Note that we use the same scale on the y- and x-axis as in the unconditional distribution above, such that the density on the left is in the order of 10^{-5} while the income is denoted in 1000s of Euros with purchasing power of the year 2010. Two aspects shall be highlighted. First, while these conditional income distributions are far more coarse with respect to the explanatory variable than the one we obtain by GAMLSS, they already give an indication for the adequacy of the various distributions we display in for each CID. Note that the black line indicates a non-parametric estimate and the pink line gives the estimates for the beta distribution of second kind. Naturally, this four parameter distribution outperforms the Dagum distribution, which we employ. But as pointed out above, it is far more difficult to estimate in a GAMLSS context. Second, is the difference of CIDs already displayed by this highly aggregated perspective.

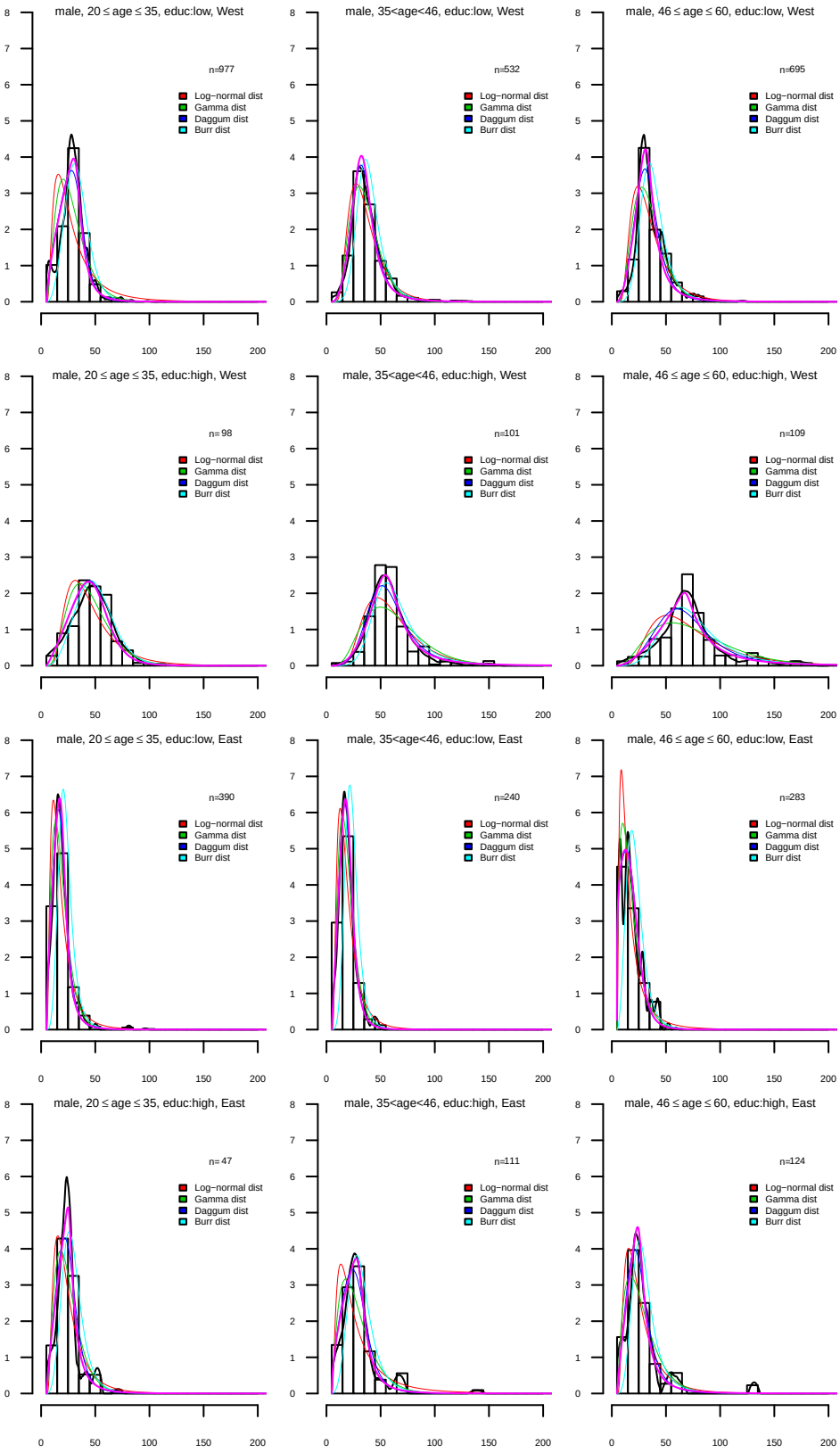

Figure 6: Income Distribution of Males in 1992

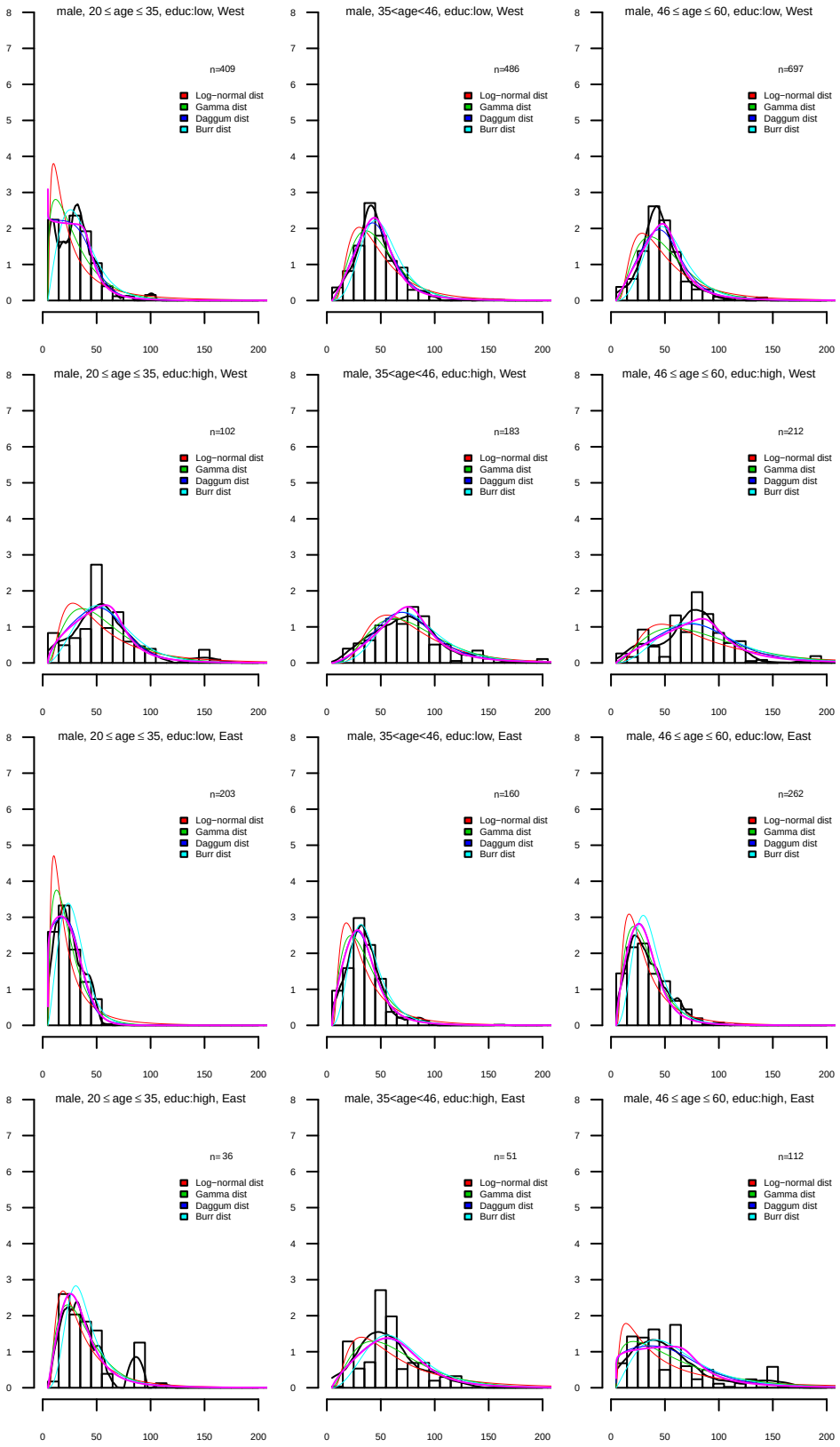

Figure 7: Income Distribution of Males in 2010

C GAMLSS

GAMLSS provides a semiparametric framework for the estimation of conditional distributions. While it relies on parametric distributions, the effect of explanatory variables on the parameter estimate entails non-parametric procedures, such as splines. For our estimation of the non-linear effects of age, we employ penalised B-splines (see Eilers and Marx, 1996). For the estimation procedure implemented in the `gamlss` package in `R` a numerical maximisation of the penalised likelihood is thus required. One aspect which must be paid particular attention is the selection of the hyperparameters for which Rigby and Stasinopoulos (2005) propose several alternatives. For our estimations we use the generalised Akaike Information criterion (see Akaike, 1983) to select an appropriate set of hyperparameters. For given hyperparameters we then maximise the penalised likelihood using the RS algorithm. For further information we refer to Rigby and Stasinopoulos (see 2005, pp.535-541).

Naturally, the estimation of whole distributions is much more involved than mean regression, as several interdependent predictors have to be estimated simultaneously. Consequently, several problems are attached to the estimation.

It must be pointed out, that as the estimation strategy relies on numerical methods it is in principle liable to the standard problems associated therewith (convergence, local maxima, etc.). Caution is thus required in the estimation process. In addition, while the Fisher information matrix of the Dagum distribution is readily available and used in the estimation process, the asymptotic standard errors obtained thereby prove inappropriate for the small sample estimation which we pursue. This problem has not been addressed so far such that momentarily we still lack satisfactory methods to reliably quantify the insecurity attached to the parameter estimates. In the frequentist setting bootstrapping methods appear most promising. Alternatively the issue of parameter uncertainty can be addressed in a Bayesian framework (see for example Klein et al., 2014).

Similarly, much work remains to be done on model selection. As Nick Longford plastically points out: "The new models are top of the range mathematical Ferraris, but the model selection that is used with them is like a sequence of tollbooths at which partially sighted operators inspect driver's licenses and road worthiness certificates." Thus models must for the moment be primarily selected on the grounds of economic reasoning.

More specifically to our estimation of Dagum distributions Kleiber and Kotz (2003, p.218) point to problems with likelihood based estimation. Domański and Jedrzejczak (1998) show in a simulation study that simple ML estimation, while consistent, is liable to biased estimation for small to moderate sample sizes. While more work must be done on this issue, preliminary simulations we conducted show that while these biases persist for small samples, their impact on the auxil-

iary measures we use is in general within an acceptable range even for very small sample sizes. Nonetheless, work on more robust estimation strategies would be necessary to make GAMLSS estimation of CID less sensitive to isolated observations. Especially the problem of the score function becoming unbounded can cause severe problems in the estimation process. Previous work by Victoria-Feser (1995) already gives a methodological framework which may be applied to account for these problems.

Despite these problems which will have to be addressed in the future, GAMLSS offers new perspectives for the analysis of conditional income distribution which we turn to now. For further discussion on GAMLSS contrasting it with mean regression, quantile regression and distribution regression see section D.

C.1 Inclusion of point masses

For modelling the whole CID, we employ two point masses next to the truncated CID. We thus partially discretize the continuous income distribution similar to Fortin and Lemieux (1998). Naturally, a better representation of the distribution is desirable but will have to be left to further research. Using these point masses the estimation of the mean of the whole CID can be simply calculated by:

$$\hat{\mu}_{CID} = \hat{p}_{pr}\hat{\mu}_{pr} + \hat{p}_{dag}\mu_{dag}, \tag{14}$$

where $\hat{p}_{pr}$ and $\hat{p}_{dag}$ are the probability of precarious incomes and the probability of incomes above 4,800€estimated by sequential logit. $\hat{\mu}_{pr}$ is the estimated mean income of those with precarious incomes. For simplicity we do not condition on age but only on education and region for the estimation of $\hat{\mu}_{pr}$. The mean of the estimated Dagum distribution μ_{dat} is given by Equation (3).

For the Theil index of each CID we use an approximation of the estimated Dagum distribution whereby we discretize it into 100,000 bins on the range 4,800€to 10,004,800€evaluating the density at the centre of each bin. We then use the standard formula for the Theil index weighting each value from the discretization as well as the mean with precarious incomes as well as zero-incomes with their corresponding density. For the latter, it should be noted that we approximate the zero-incomes by a symbolic cent, as the Theil index is only defined for positive incomes. For the Gini coefficient, which we also calculated, but didn't display, we proceeded analogously.

For the quantile ranges and the shares above the threshold values we can simply incorporate the point masses into the distribution and determine the quantile and p-values for the desired level. It should be noted that the point masses for the precarious incomes cause the slight discontinuities for the interquantile ranges observed in Figure 3 and 5.

C.2 Distributions considered for CIDs

Next to the Dagum distribution which we chose as the most appropriate distribution, we also fitted various other distributions using GAMLSS[24]:

- The log-normal distribution is probably the most popular distribution for modelling conditional income distributions. Yet our results show that at least for the conditional income distributions we considered, the log-normal distribution proved inappropriate as can be seen in Section E.2.
- As we found varying kurtosis among the conditional distribution of log-incomes, we also experimented with the exponential power distribution, also known as the generalised normal distribution. Yet we found that the constraints of a symmetric structure proved inadequate for log-transformed incomes.
- The gamma distribution was considered and in comparison to the log-normal it generally had a better performance, yet performed considerably worse than the three-parameter Sing-Maddala and Dagum distribution.
- Brachmann et al. (1996) find the Sing-Maddala distribution to adequately describe the aggregate income distribution from 1984 to 1993. Yet for CIDs our results found its fit to be worse than for the Dagum distribution.

As the choice of distribution is pivotal to the modelling of CIDs by GAMLSS much more work must be done on this issue though.

[24]For a list of the wide variety of distributions which have been implemented see Stasinopoulos and Rigby (2007) or the GAMLSS website www.gamlss.org.

D Other regression approaches

D.1 Mean regression

Standard analysis with regard to conditional incomes focusses on the conditional mean and generally treats additional parameters as nuisance parameters. While additional aspects of the conditional income distribution like heteroskedasticity or varying shapes can be partially incorporated by the use of link functions, the flexibility remains constrained by the variability of only one predictor. Nonetheless, using GLM or GAM conditional income distributions of the exponential family (e.g. log-normal or gamma) can in principle be modelled. For reasons of comparison we fit an generalised additive model of the same form as for our GAMLSS equations:

$$\log(y) = s_1(age) + Hs_2(age) + Es_3(age) + HEs_4(age) + \varepsilon, \tag{15}$$

where y denotes the truncated income-vector excluding all incomes up to 4,800€ and where the elements of ε are assumed to be independently distributed following a distribution from the exponential family, e.g. the normal distribution centred around zero with a constant variance σ^2. The thusly estimated means and Theil indices for the whole CIDs are displayed in Figure 8. We can observe that the expected incomes are not obviously worse than in the GAMLSS approach. Nonetheless, aspects wheref scale and shape are important, like for the Theil index, we can see that the log-normal distribution misfits and thus portrays much poorer estimates for the Theil indices of CIDs.

D.2 Mixed models

The inclusion of random effects in the distribution allows for random intercepts and/or random slopes for different groups. Yet one core requirement for modelling CIDs is the ability to use continuous variables, like age as explanatory factors influenceing the distribution. As random effects can only be assigned to a finite number of groups and thus not in accordance with a continuous variable mixed models are not considered in detail here.

D.3 Distribution regression

A third alternative for the modelling of CIDs is the new class of so called distribution regressions (see Chernozhukov et al., 2013, p.10). It stems from the literature of survival functions and relies on link functions to model the conditional cumulative distribution function. Further in-depth

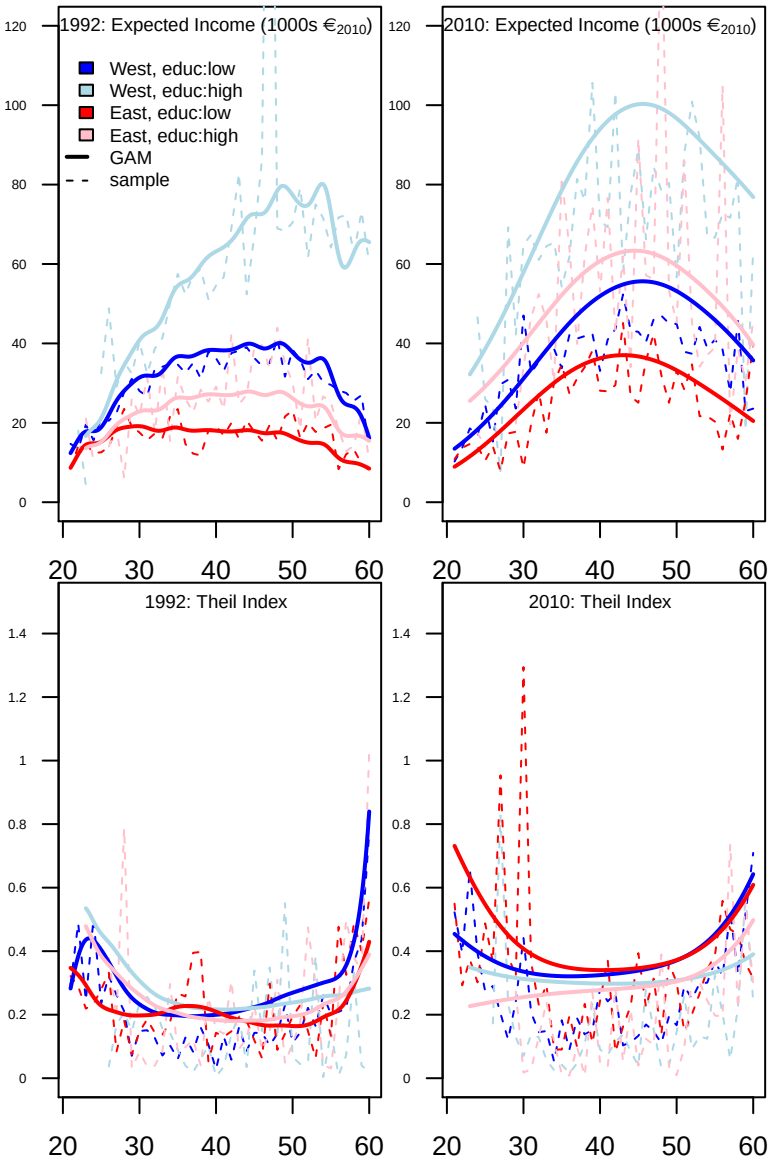

Figure 8: Means and Theil Indices of whole CID obtained by GAM

comparisons with this model class is needed. Nonetheless, in principle the flexibility of the CID is constrained as only one predictor is used, similarly to the case of mean regression.

D.4 Quantile Regression

Quantile Regression, pioneered by Koenker and Bassett (1978), minimises the check function and thus allows for the estimation of a conditional quantile function. In principle we are thus able to estimate any quantile τ of the CID:

$$y = s_{1,\tau}(age) + Hs_{2,\tau}(age) + Es_{3a,\tau}(age) + HEs_{4,\tau}(age) + \varepsilon, \tag{16}$$

While this alternative, which has been used for the analysis of income distributions by Machado and Mata (2005), offers a non-parametric alternative to the estimation of CIDs, some possible problems with such an approach should be pointed out.

It is the nature of quantile distribution that it is intrinsically focussed towards the estimation of specific quantiles within the conditional distribution rather than the conditional distribution at large. While for some purposes this bears several advantages (for example interquantile ranges are estimated with greater ease and less assumptions) it can prove a disadvantage if a comprehensive analysis of the CIDs is required. If enough data is available the problem is solely of computational nature as a large number of conditional quantiles can be estimated, constructing the conditional distribution thereof. Yet, as is generally the case, the scarcity of data is a major concern in which case quantile regression as well as any non-parametric approximation becomes highly unstable. The constraints that a parametric approach imposes can in that case aide to get a better approximation of the CID, if and only if the imposed constraints are applicable.

Connected to this aspect, is the role of censored data. As Bach et al. (2009) point out, the SOEP under-represents the top incomes. While it is in principle possible to account for this censored data in non-parametric approaches as well, the same argument as before applies, that a parametric approach can lend structure to the appropriate modelling of censored data. While the non-parametric approach employing quantile regression thus offers several advantages of GAMLSS, most notably the greater flexibility in modelling the CID, this flexibility can especially for small sample sizes hamper the estimation process as it fails to lend the required stability. Whether this additional flexibility provided by our parametric approach is of more use than harm, hinges on the question of whether an appropriate parametric form for the CIDs can be found.

E The truncated CID

E.1 Auxiliary measures for interpretation of the truncated CID

Figures 9 and 10 display several standard measures for size distributions, namely the expectation of the conditional distribution, its standard deviation and skewness as well as the Theil Index which incorporates all three moments and is a well known measure for inequality. The thick lines display the resulting measure from the GAMLSS estimation of the conditional distribution, while the dashed lines display the measure obtained directly from the sample for a given age, education level and region. It is well noted in the literature that the sample measures for moments are biased, especially for samples resulting from distributions other than the normal (Joanes and Gill, 1998, see). While we have accounting for this sampling bias in the standard manner[25] it is likely that some bias remains. Interestingly, in 1992 the skill premium in East Germany is relatively small such that on average incomes of men with a degree in the East are still lower than those of men without higher education in the West. The distributions' standard deviations show higher dispersion with higher age and rising mean incomes. We can also see that from the late thirties onwards the dispersion of men from the East with higher education exceed those without higher education in the West. Even more striking are differences in the skewness, which show that especially in later years the higher spread is anything but symmetric but largely caused by some high incomes. In comparison the skewness of the other distributions seems negligible. Nonetheless, we can observe a similar trend as for the standard deviation, i.e. increasing skewness as age and incomes progress. It should also be noted that skewness is systematically higher in the East for both education levels. This indicates that in the East more incomes are notably clustered at the lower end of the income range.The resultant findings for the Theil index for these CIDs show that the thus measured inequality is greater in the East than in the West, a finding which is contrary to the popular perception. However, as pointed out above, the nature of this higher inequality is such that a large share of incomes is clustered within the lower range of incomes with few very high incomes. While the perception of the income distribution which is probably mainly driven by the dispersion of the inner quantiles (e.g. 15^{th}-85^{th}), i.e. that of the 'ordinary man', the nature of income inequality in the East is very different to that of the West (see below) as well as the inter-group differences, which we do not discuss in detail here.

However, the truncated distribution we have considered so far is only telling part of the story. As a matter of fact, much of the literature on income inequality is hampered by such a partial perspective, where only parts of the population (only males, full-time employees, etc.) and/or only

[25]For the standard deviation we have used the adjustment proposed by the cov.wt function from the stats package in R. For the skewness we have used the sample adjustment which is found for weighted data in the statistical software SAS. For the Theil index we have refrained from correcting the bias, following the ineq package in R.

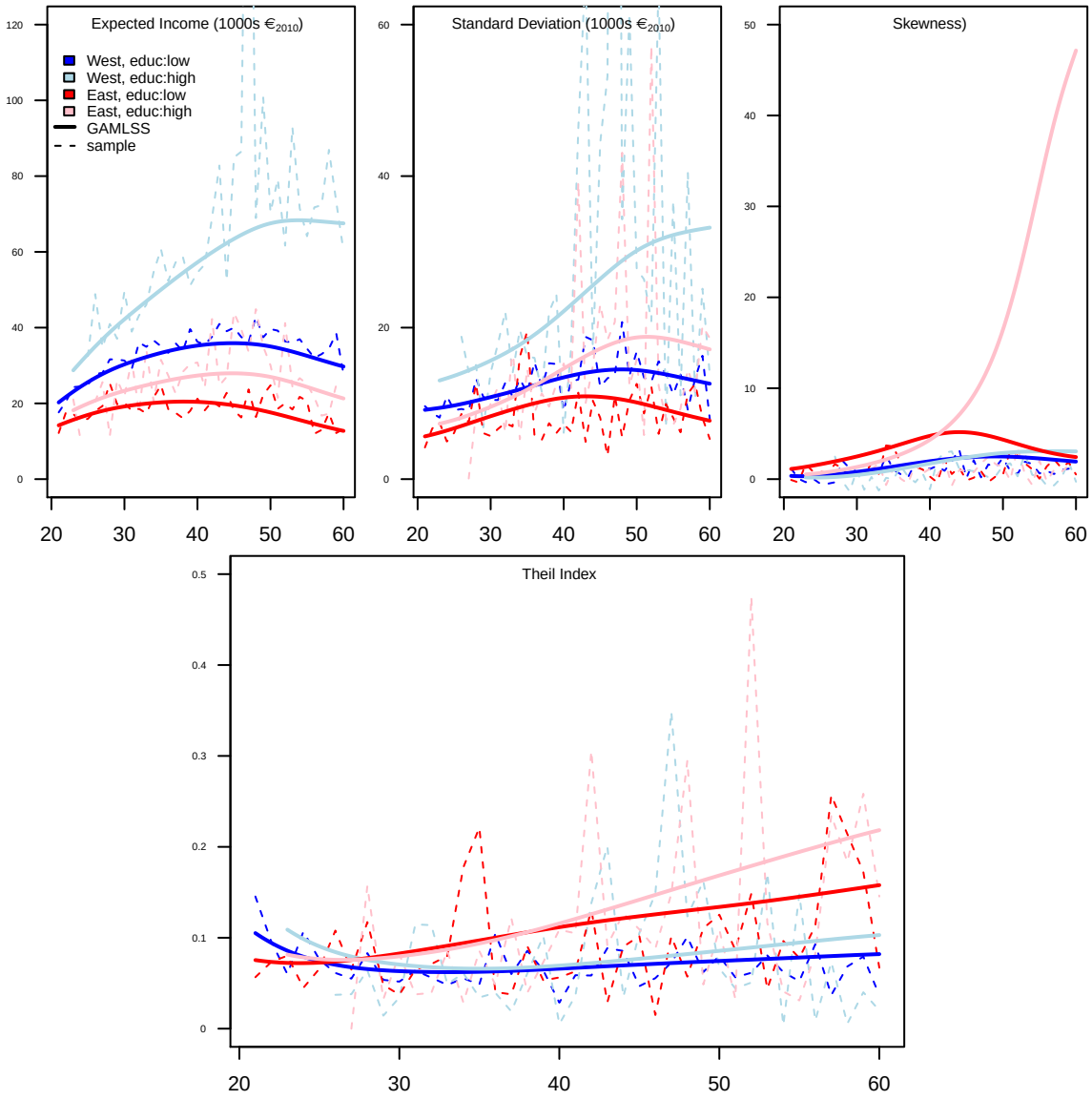

Figure 9: Measures for truncated CIDs males in 1992

parts of the income range (only incomes of people in employment, i.e. those above zero). While such partial perspectives provide scientific insight on their own a comprehensive account of income inequality must consider the whole income distribution, despite the analytical problems associated with such an approach.

The most noticeable difference of the expectation for the dependent income distributions displayed on the top left of Figure 10 is the surpassing of the average incomes of men with higher education in the East of those without higher education in the West. Also the rift of incomes between East and West for men without higher education has also narrowed. This convergence in mean incomes already pointed out by Vollmer et al. (2013) and others. Especially for the older generations in the East the increase in income dispersion is dramatic. For those with higher education this is

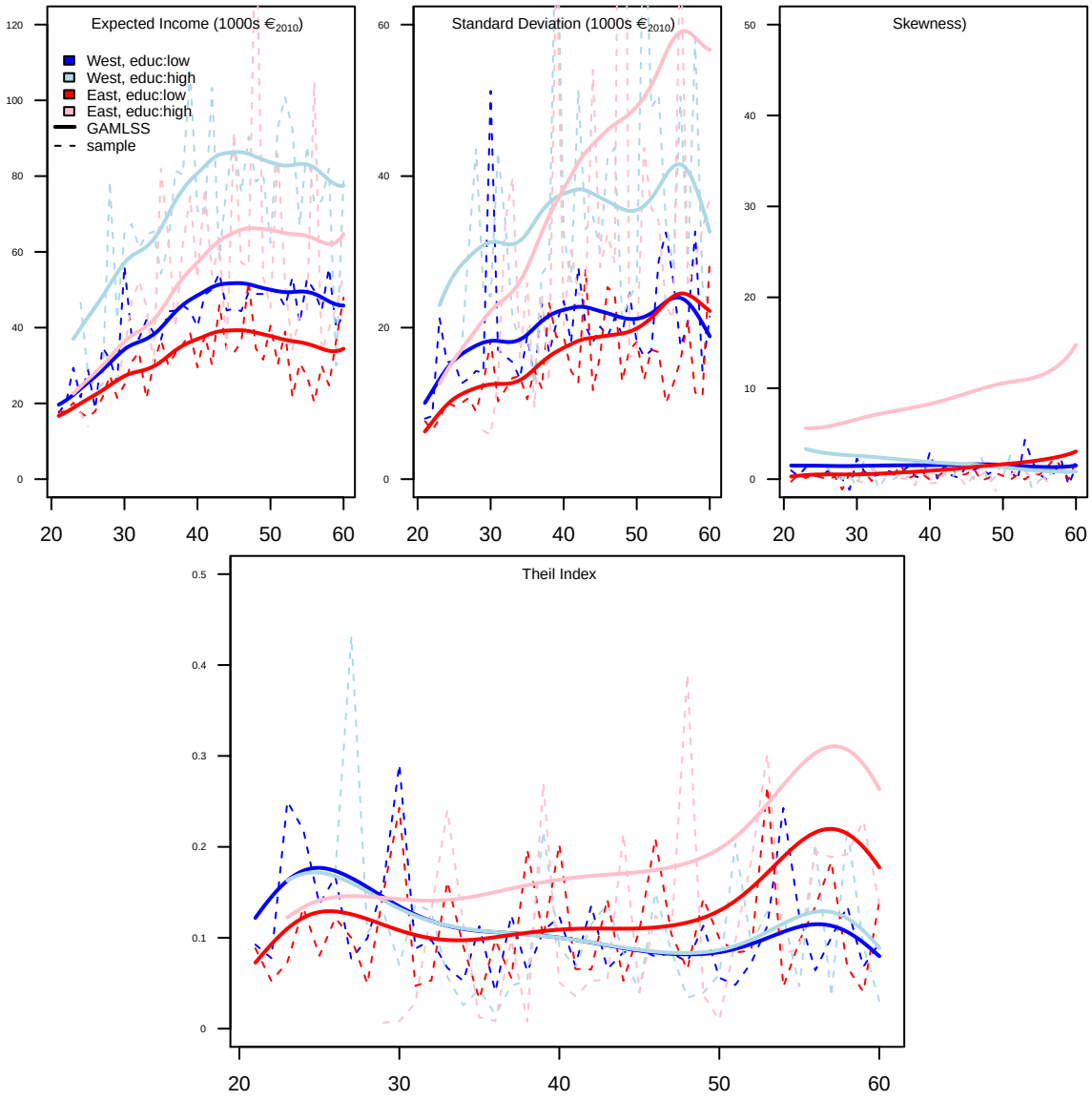

Figure 10: Measures for truncated CIDs of males in 2010

combined with a persistent high skewness again causing this group to show high within-inequality as indicated by the Theil index. With respect to the convergence of the whole income distribution (rather than just the mean) one has to concede that true convergence is still some way off. For the truncated income distribution, mean convergence across the age groups is driven by the catching up of incomes for some with a large part of the male population in the East left wanting as indicated by the higher standard deviation and skewness.

E.2 Testing the quality of fit for the truncated CID

In order to test whether the Dagum distributions as specified by our GAMLSS estimation adequately model the data, we test the hypothesis that

$$H_0 : f(x) = f_0(x, \theta), \tag{17}$$

where $f(x)$ is the observation generating p.d.f. and $f_0(x, \theta)$ is the parametric distribution thought to model the data.

In their paper for the assessment of parametric income distributions Brachmann et al. (1996) propose the test by Bickel and Rosenblatt (1973). Yet one of the problems with this test is that it is heavily influenced by the bandwidth selection. Especially, the expected non-parametric distribution under H_0 given in the equation below

$$E_0\big(\hat{f}(x)\big) = \frac{h^2}{2} f_0''(x, \hat{\theta})\mu_2(K) + f_0(x, \hat{\theta}) + o(h^2), \quad \text{with } \mu_2(K) = \int u^2 K(u) du \tag{18}$$

poses problems for large h. As due to the small sample sizes we often encounter and the resultant large bandwidths proposed by standard methodology (e.g. Sheather and Jones (1991)) this test was found to be not suited for our purpose. In its stead we use a bootstrapped version of the Kolmogorov-Smirnov Test, which use Monte Carlo simulations to obtain the distribution of the test-statistic. The test statistic is given by (see for example Hartung et al., 1987, p.183)

$$D_n = \sup_x \mid F_0(x, \theta) - S_n(x) \mid, \tag{19}$$

where $F_0(x, \theta)$ denotes the c.d.f. of our parametric fit from Equation (2) and $S_n(x)$ denotes the empirical cumulative distribution function for observations $x_1, \ldots, x_n$, with n being the sample size of the given subpopulation under consideration. The distribution of this test-statistic was then obtained by parametric bootstrap whereby we used 100,000 simulations samples of size n for each subpopulation yielding a distribution of the test-statistic. Using this procedure we obtained the p-values for each subpopulation given in Table 1 and 2.

For a p-value we expect to see a 5% share of observations to show a test-statistic with a corresponding p-value greater than 0.05. On average over both time periods we get an average rejection rate of 0.069, which is just above the 0.05 we would expect. While it must be noted that for the males without higher education in the West in 1992 we generally have slightly too high shares of rejections, they do portray that there is no systematic problems with the modelling of these income distributions by the Dagum distribution. We therefore conclude that GAMLSS employing the Dagum distribution adequately models the CIDs under consideration.

Age	LowEduc.West	HighEduc.West	LowEduc.East	HighEduc.East
21	0.100		0.113	
22	0.522	0.606	0.011	
23	0.039		0.545	0.693
24	0.906		0.283	0.265
25	0.163	0.961	0.816	
26	0.105	0.464	0.772	0.985
27	0.037	0.854	0.456	0.166
28	0.363	0.732	0.002	0.189
29	0.166	0.266	0.269	0.135
30	0.055	0.148	0.397	0.777
31	0.445	0.925	0.509	0.740
32	0.336	0.756	0.260	0.318
33	0.904	0.537	0.326	0.218
34	0.027	0.409	0.807	0.239
35	0.758	0.015	0.498	0.634
36	0.641	0.823	0.372	0.688
37	0.147	0.398	0.117	0.747
38	0.757	0.290	0.021	0.559
39	0.387	0.814	0.170	0.569
40	0.211	0.342	0.185	0.810
41	0.610	0.598	0.557	0.705
42	0.832	0.957	0.884	0.898
43	0.505	0.031	0.306	0.631
44	0.302	0.651	0.159	0.498
45	0.543	0.548	0.909	0.203
46	0.058	0.620	0.289	0.472
47	0.344	0.432	0.343	0.678
48	0.066	0.885	0.580	0.135
49	0.481	0.256	0.056	0.014
50	0.477	0.337	0.093	0.530
51	0.950	0.164	0.030	0.416
52	0.178	0.696	0.056	0.051
53	0.860	0.166	0.017	0.242
54	0.100	0.051	0.033	0.037
55	0.156	0.856	0.029	0.092
56	0.530	0.389	0.134	0.921
57	0.210	0.126	0.163	0.864
58	0.894	0.013	0.576	0.479
59	0.057	0.310	0.043	0.700
60	0.540	0.604	0.752	0.849
Share of $p<0.05$	0.075	0.075	0.200	0.050

Table 1: Dagum: P-values from KS-Test for Males 1992

Age	LowEduc.West	HighEduc.West	LowEduc.East	HighEduc.East
21	0.127		0.934	
22	0.759		0.970	
23	0.300		0.553	
24	0.150	0.114	0.934	0.647
25	0.543	0.729	0.366	0.331
26	0.361	0.082	0.965	
27	0.130	0.659	0.935	0.822
28	0.159	0.744	0.134	
29	0.341	0.561	0.278	0.790
30	0.236	0.714	0.424	0.680
31	0.552	0.417	0.263	0.807
32	0.794	0.960	0.821	0.539
33	0.275	0.393	0.482	0.408
34	0.208	0.912	0.261	0.080
35	0.756	0.352	0.543	0.185
36	0.113	0.538	0.011	0.006
37	0.563	0.713	0.518	0.792
38	0.437	0.407	0.762	0.639
39	0.080	0.968	0.266	0.670
40	0.162	0.638	0.326	0.639
41	0.821	0.238	0.976	0.660
42	0.532	0.592	0.077	0.762
43	0.772	0.435	0.703	0.292
44	0.314	0.328	0.264	0.163
45	0.152	0.193	0.380	0.344
46	0.668	0.784	0.582	0.995
47	0.743	0.595	0.918	0.929
48	0.972	0.580	0.237	0.367
49	0.444	0.736	0.011	0.197
50	0.671	0.430	0.657	0.866
51	0.385	0.788	0.534	0.545
52	0.874	0.848	0.060	0.991
53	0.460	0.846	0.010	0.905
54	0.872	0.858	0.220	0.853
55	0.399	0.508	0.036	0.605
56	0.402	0.730	0.202	0.999
57	0.756	0.385	0.094	0.309
58	0.203	0.852	0.554	0.984
59	0.274	0.405	0.048	0.209
60	0.342	0.764	0.213	0.642
Share of $p<0.05$	0.000	0.000	0.125	0.025

Table 2: Dagum: P-values from KS-Test for Males 2010

Age	LowEduc.West	HighEduc.West	LowEduc.East	HighEduc.East
21	0.100		0.113	
22	0.522	0.606	0.011	
23	0.039		0.545	0.693
24	0.906		0.283	0.265
25	0.163	0.961	0.816	
26	0.105	0.464	0.772	0.985
27	0.037	0.854	0.456	0.166
28	0.363	0.732	0.002	0.189
29	0.166	0.266	0.269	0.135
30	0.055	0.148	0.397	0.777
31	0.445	0.925	0.509	0.740
32	0.336	0.756	0.260	0.318
33	0.904	0.537	0.326	0.218
34	0.027	0.409	0.807	0.239
35	0.758	0.015	0.498	0.634
36	0.641	0.823	0.372	0.688
37	0.147	0.398	0.117	0.747
38	0.757	0.290	0.021	0.559
39	0.387	0.814	0.170	0.569
40	0.211	0.342	0.185	0.810
41	0.610	0.598	0.557	0.705
42	0.832	0.957	0.884	0.898
43	0.505	0.031	0.306	0.631
44	0.302	0.651	0.159	0.498
45	0.543	0.548	0.909	0.203
46	0.058	0.620	0.289	0.472
47	0.344	0.432	0.343	0.678
48	0.066	0.885	0.580	0.135
49	0.481	0.256	0.056	0.014
50	0.477	0.337	0.093	0.530
51	0.950	0.164	0.030	0.416
52	0.178	0.696	0.056	0.051
53	0.860	0.166	0.017	0.242
54	0.100	0.051	0.033	0.037
55	0.156	0.856	0.029	0.092
56	0.530	0.389	0.134	0.921
57	0.210	0.126	0.163	0.864
58	0.894	0.013	0.576	0.479
59	0.057	0.310	0.043	0.700
60	0.540	0.604	0.752	0.849
Share of p<0.05	0.075	0.075	0.200	0.050

Table 3: Log-normal: P-values from KS-Test for Males 1992

Age	LowEduc.West	HighEduc.West	LowEduc.East	HighEduc.East
21	0.000		0.000	
22	0.000		0.000	
23	0.000		0.000	
24	0.000	0.990	0.000	0.992
25	0.000	1.000	0.000	0.978
26	0.000	0.000	0.000	
27	0.000	0.998	0.000	0.994
28	0.000	0.000	0.000	
29	0.000	0.000	0.000	1.000
30	0.000	0.000	0.000	0.000
31	0.000	0.000	0.000	0.000
32	0.000	0.000	0.000	0.000
33	0.000	0.000	0.000	0.000
34	0.000	0.000	0.000	0.997
35	0.000	0.000	0.000	0.865
36	0.000	0.000	0.000	0.947
37	0.000	0.000	0.000	0.000
38	0.000	0.000	0.000	0.994
39	0.000	0.000	0.000	0.000
40	0.000	0.000	0.000	0.999
41	0.000	0.000	0.000	1.000
42	0.000	0.000	0.000	1.000
43	0.000	0.000	0.000	0.000
44	0.000	0.000	0.000	0.984
45	0.000	0.000	0.000	0.000
46	0.000	0.000	0.000	0.000
47	0.000	0.000	0.000	0.000
48	0.000	0.000	0.000	0.000
49	0.000	0.000	0.000	0.000
50	0.000	0.000	0.000	1.000
51	0.000	0.000	0.000	0.000
52	0.000	0.000	0.000	0.000
53	0.000	0.000	0.000	0.000
54	0.000	0.000	0.000	0.000
55	0.000	0.000	0.000	0.000
56	0.000	0.000	0.000	0.000
57	0.000	0.000	0.000	0.000
58	0.000	0.000	0.000	0.000
59	0.000	0.000	0.000	0.000
60	0.000	0.000	0.000	0.000
Share of p<0.05	1.000	0.850	1.000	0.550

Table 4: Log-normal: P-values from KS-Test for Males 2010